

Copy No. _____



Pass-fail performance testing for detection systems

Ronald T. Kessel

Defence R&D Canada

Technical Memorandum

DREA TM 2001-205

February 2002



National
Defence

Défense
nationale

Canada



Pass-fail performance testing for detection systems

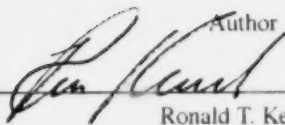
Ronald T. Kessel

Defence Research Establishment Atlantic

Technical Memorandum

DREA TM 2001-205

February 2002

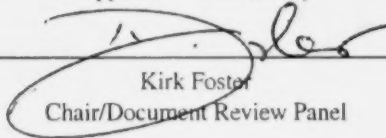
Author


Ronald T. Kessel

Approved by


Dr. Ron Kuwahara
Head/EM Section

Approved for release by



Kirk Foster
Chair/Document Review Panel

Abstract

Detection is an uncertain operation subject to many random factors. The performance of a detection system is therefore specified probabilistically, by way of its probabilities of detection and false alarm, and the evaluation of a system's performance falls, unavoidably, within the scope of probability and statistics. In military applications, the central role of probability and statistics has too often been upstaged by the novelty of new detection technology, which, to demonstrate in all of its features, typically leaves little time or inclination for a detailed treatment of performance probabilities. But a detailed statistical analysis performance is crucial for drawing objective conclusions from a performance test, such as whether the system passes or fails minimum operational requirements. The statistics of performance testing are reviewed here, as a means to manage the measurement uncertainties in pass-fail system testing. Two different decision methods are presented, hypothesis testing and Bayesian inference, each with their particular approach to manage uncertainties, yet both working toward the same end. Pass-fail judgements drawn from "perfect" test results (no missed targets, and no false alarms) are given special consideration because they are often encountered in practice owing to small sample sizes. The minimum number of dummy targets required for a performance test is derived and serves as a rough guide when planning and evaluating performance demonstrations.

Résumé

La détection est une opération incertaine qui est soumise à de nombreux facteurs aléatoires. Les performances d'un système de détection doivent donc être établies de façon probabiliste, c'est-à-dire en tenant compte des probabilités de détection et des fausses alarmes; l'évaluation de la performance d'un système relève inévitablement du domaine des probabilités et des statistiques. Dans les applications militaires, le rôle central que jouent les probabilités et les statistiques a trop souvent été relégué au second plan par l'aspect novateur des nouvelles technologies de détection, qui, généralement, ne laissent guère de temps pour effectuer un calcul détaillé des probabilités de performance, lequel permettrait d'en démontrer toutes les caractéristiques, et ne suscitent guère d'intérêt à cet égard. Il est cependant crucial de procéder à une analyse statistique afin de tirer des conclusions objectives d'un essai de performance permettant, entre autres, de déterminer si un système répond ou non à des exigences opérationnelles minimales. Dans l'article, on analyse les statistiques des essais de performance, afin de gérer l'incertitude des mesures faites lors d'essais du système. On présente deux méthodes de décision différentes, soit la vérification des hypothèses et l'inférence bayésienne. Chaque méthode offre une approche particulière pour gérer les incertitudes tout en cherchant à atteindre le même objectif. On examine les jugements sur la réussite ou l'échec fondés sur des résultats $\hat{z}$ parfaits (aucune cible manquée et aucune fausse alarme), car de tels jugements sont fréquents en

pratique, en raison de la petite taille des échantillons. On calcule le nombre minimal de cibles factices nécessaires pour effectuer un essai de performance, afin de fournir un chiffre approximatif lors de la planification et de l'évaluation des démonstrations de la performance.

Executive summary

Background

Proven performance is crucial for military detection systems. A performance test usually entails the deployment and subsequent search for dummy targets under realistic operating conditions. Such tests should be a central part of the design, evaluation, and procurement of any detection system. But detection is an uncertain operation. If the conclusions drawn about performance from the test are to be objective and justified, then the inherent uncertainties of the performance test must be managed in an objective, unbiased way.

Principal results

The uncertainties of pass-fail testing are reviewed here, and two important questions are addressed: 1) when planning a test, how many dummy targets should be deployed to prove that a detection system exceeds a given performance level?; and 2) when assessing a test, with what confidence can an objective observer claim that the detection system passes or fails its performance requirements? Special attention is given furthermore to "perfect" test results (no missed targets, and no false alarms), which are often encountered in practice and in the literature owing to small sample sizes—the objective being again to decide whether the system passes on the basis of that perfect result, or whether further testing is required.

Significance of results

A clear understanding of the statistics behind performance testing is necessary both when writing system performance specifications, and when evaluating a system against those specifications. The central role of probability and statistics has nevertheless been overlooked, no doubt because the novelty of new system technology, and the effort required to demonstrate its new features, leaves little time or inclination for a comparatively mundane, but nonetheless important, discussion of performance probabilities and statistics. The need for proven system performance comes now to the forefront as automated detection systems evolve towards greater operational use.

The present analysis lays the groundwork for performance testing in the development of automated detection aids in the Remote Minehunting System (RMS) Technology Demonstrator (TD) project now underway at DREA, but it can also be applied much more widely, for the evaluation of other detection systems, whether military, industrial, or medical.

Future Work

The usual single-point specification of performance, by way of a minimum probability of detection and maximum probability of false alarm, will be extended to a more

general, minimal receiver-operator characteristic (ROC) curve, with a view to developing new pass-fail specifications that allow for a flexible range of detection and false alarm performance levels, any of which may be operationally preferred, depending on variable operational factors, such as changing background clutter levels, known risk, and so forth.

Ronald T. Kessel. 2002. Pass-fail performance testing for detection systems.
DREA TM 2001-205. Defence Research Establishment Atlantic.

Sommaire

Contexte

La performance éprouvée est un aspect crucial des systèmes de détection militaires. Un essai de performance comprend en général le déploiement et la recherche ultérieure de cibles factices dans des conditions opérationnelles réalistes. Ces essais devraient constituer un élément essentiel de la conception, de l'évaluation et de l'acquisition d'un système de détection. Cependant, la détection est une opération incertaine. Il faut gérer de façon objective et impartiale l'incertitude inhérente aux essais de performance si l'on veut tirer des conclusions objectives et justifiées sur la performance d'un système.

Principaux résultats

On analyse dans cet article les incertitudes liées aux essais visant à déterminer si un système répond ou non à des exigences opérationnelles minimales, c.-à-d. réussite/échec, et on pose deux questions d'importance : 1) Lors de la planification, combien de cibles factices doit-on déployer pour prouver qu'un système de détection dépasse un niveau de performance donné? et 2) Lors de l'évaluation d'un essai, avec quelle certitude un observateur objectif peut-il affirmer qu'un système de détection répond ou non aux spécifications de performance minimales? On accorde en outre une attention particulière aux résultats à parfaits (aucune cible manquée et aucune fausse alarme), que l'on observe fréquemment en pratique et dans la littérature en raison de la petite taille des échantillons l'objectif étant, encore ici, de déterminer, à partir de ce résultat parfait, si le système répond ou non aux spécifications minimales ou s'il est nécessaire de procéder à d'autres essais.

Importance des résultats

Il faut bien comprendre les statistiques derrière les essais de performance lorsqu'on établit les spécifications relatives à la performance et lorsqu'on évalue un système en fonction de ces spécifications. Néanmoins, on néglige de tenir compte du rôle central que jouent les probabilités et les statistiques, sans doute parce que l'aspect novateur des nouvelles technologies et l'effort qu'exigerait la démonstration de leurs nouvelles caractéristiques ne laissent guère de temps et suscitent peu d'intérêt pour une discussion, comparativement terre à terre mais néanmoins importante, sur le rôle des probabilités et des statistiques dans la détermination de la performance. La performance éprouvée des systèmes se retrouve maintenant au premier plan, avec l'utilisation de plus en plus fréquente de systèmes de détection automatiques lors d'opérations. La présente analyse établit la base à partir de laquelle seront élaborés les essais de performance qui serviront lors de la mise au point d'aides à la détection automatique dans le cadre du projet de démonstration de la technologie du système télécommandé de chasse aux mines, présentement en cours au CRDA. Cette analyse peut également servir à évaluer beaucoup d'autres systèmes de détection utilisés à des fins militaires, industrielles ou médicales.

Travaux prévus

Les spécifications de performance à point unique, qui sont habituellement établies en fonction de la probabilité minimale de détection et de la probabilité maximale de fausse alarme, seront appliquées à une courbe minimale plus générale illustrant les caractéristiques du récepteur/opérateur, en vue d'établir de nouvelles spécifications minimales (réussite/échec) qui permettront de choisir, selon l'opération, une portée de détection et des niveaux de performance relatifs aux fausses alarmes qui soient plus flexibles, selon les facteurs opérationnels variables, comme le changement des niveaux de fouillis d'échos, les risques connus, etc.

Ronald T. Kessel. 2002. Essais de performance échec/réussite pour les systèmes de détection. DREA TM 2001-205. Centre pour la Recherche de la Défense Atlantique.

Table of contents

Abstract	i
Résumé	i
Executive summary	iii
Sommaire	v
Table of contents	vii
List of figures	viii
1. Introduction	1
2. Preliminary remarks about detection performance testing	2
2.1 Demonstration of uncertainties	2
2.2 Single-scenario test	3
2.3 Clutter events	4
2.4 Performance specifications	4
2.5 The marginal detection system	5
2.6 Measured performance: a binomial random variable	6
3. Uncertainty analysis	7
3.1 Confidence intervals	7
4. Hypothesis testing	9
4.1 Interpreting the perfect result	14
5. Bayesian inference	16
5.1 Interpreting the perfect result	18
6. Conclusions	21
6.1 Future work	22
References	23

List of figures

- 1 A demonstration of pass-fail test uncertainties: The points represent 300 independent pass-fail performance tests of a given detection system (dashed curve), relative to minimum performance specifications (solid curve), using 10 dummy mines and 100 clutter events in each test. 3

- 2 Three confidence intervals are plotted here as a function of the measured performance, assuming that the sample size is 20. If the measured probability of detection were $P_1 = 0.80$ (on the horizontal axis), for instance, then, using the solid lines, the probability that the true system performance lies between 0.65 and 0.95 (vertical axis) is greater than 90 %. The three intervals show how the interval widens as the confidence probability is increased (dashed lines relative to solid), but narrows as the sample size is increased (dotted lines relative to solid). The vertical lines near the extreme edges of the graph mark the points at which the asymmetry in the binomial distribution makes central intervals impossible, as noted in an earlier footnote. 9

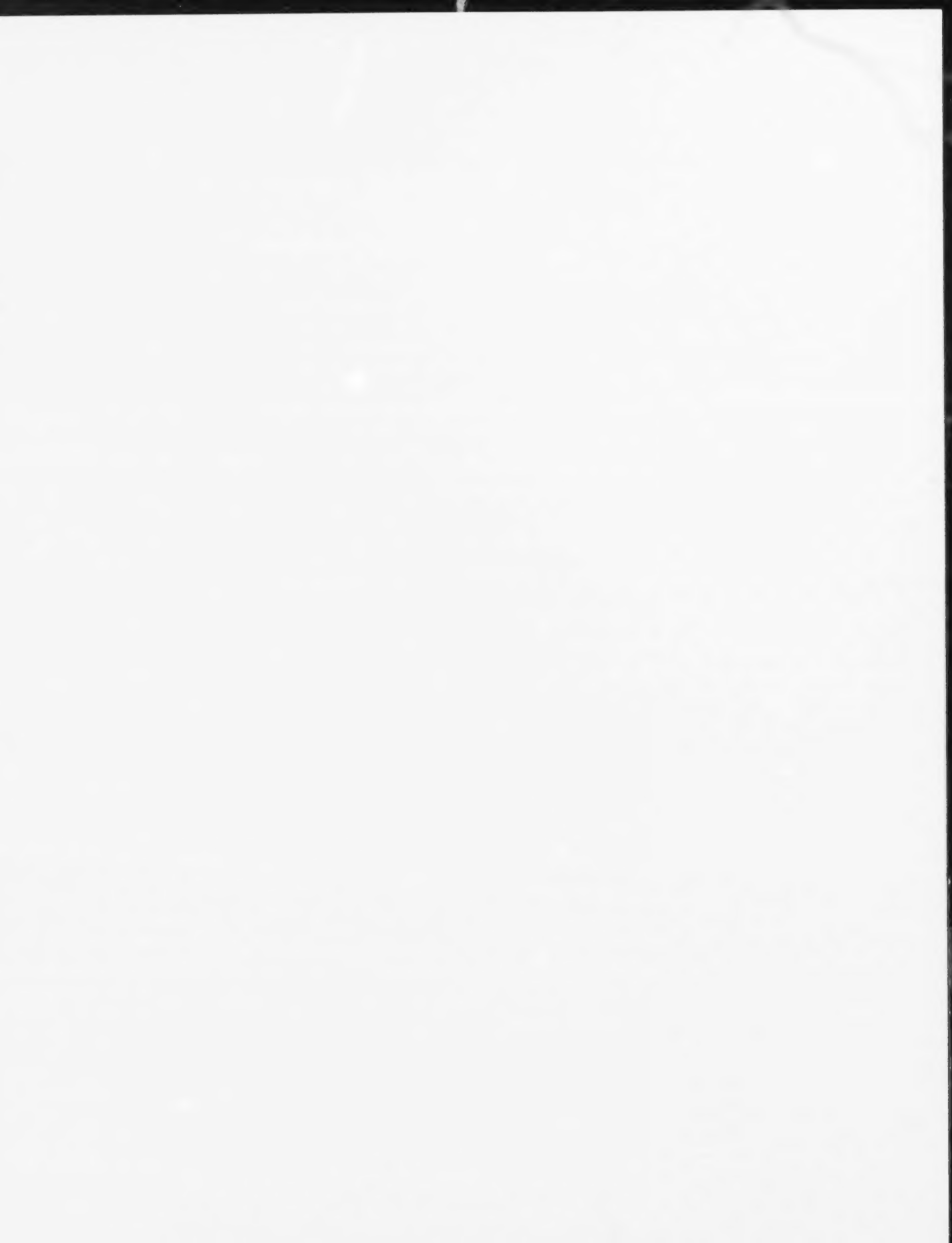
- 3 The pass-fail decision thresholds can be determined graphically. A detection (class 1) performance test is assumed here. The sample size for the figure is $n_1 = 20$, and the performance specification is $s_1 = 0.80$. The horizontal axis represents trial threshold values, and the vertical axis is the probability of making an upper or lower TYPE 1 error. The threshold values at which these equal $\alpha/2$ are the pass-fail threshold values for the test. If the measured performance falls below t_1^A , then the system fails to meet the detection specification s_1 . If it falls above t_1^B , then the system passes. If it falls between the two thresholds, then more testing (a larger sample size) is required. 12

- 4 The pass-fail decision boundaries in Figure (1) can also be determined numerically. This has been done here for a range of detection performance specifications s_1 (horizontal axis), and sample sizes $n_1 = 10, 20$ and 50 , assuming $\alpha_1 = 0.10$ 13

- 5 The minimum number of dummy targets for a performance test presumes, very optimistically, that all of the targets deployed in a random, unbiased test will be detected. In practice, significantly more targets would be required. 15

- 6 The pass probability is plotted here as a function of the measured detection performance, assuming sample sizes of 10, 20, and 50 dummy targets. The horizontal dashed line marks the $c_1 = 0.90$ pass probability. The system under test passes with regards to detection if the bar associated with the value of the performance measured in the test exceeds this level. 18

- 7 The system passes with respect to detection when the measured detection performance exceeds the PASS decision threshold, plotted here as a function of the performance requirement s_1 , assuming that there are $n_1 = 20$ dummy targets in the test. Ever more stringent thresholds result as the pass probability c_1 is increased. 19
- 8 The system passes with respect to detection when the measured detection performance exceeds the PASS decision threshold, plotted here as a function of the performance requirement s_1 , assuming that the pass probability $c_1 = 0.90$, for three different sample sizes, $n_1 = 10, 20$ and 50 dummy targets. The pass threshold approaches the diagonal as the sample size increases. 20



1. Introduction

Proven performance ranks among the most important properties of automated detection systems for military applications. For instance, a new computer-aided detection (CAD) system must exceed minimal performance specifications to be accepted for operational use, prospective detection systems may be ranked according to their performance in realistic trials carried out when purchasing a new detection system, or, in much the same way, prospective automatic target recognition (ATR) algorithms might be ranked according to their performance against a set of target and clutter images. In all such cases, detection performance must be assessed quantitatively and objectively.

A performance test entails the deployment of a number of dummy targets under realistic conditions, and then a search for those targets using the detection system under test. The percentage of targets correctly detected is an estimate of the probability of detection, and the number of false alarms is indicative of the probability of false alarm. What must not be overlooked, however, are the uncertainties inherent in the experimental method. For as with any experimental measurement, to merely report a final measured value, without explicit or implied uncertainty bounds, is to provide only half of the information that a decision maker must have, because it says nothing about the reliability of the measured result. Uncertainties determine the confidence that may be placed in the conclusions drawn from a measurement. They distinguish careful precision from a wild guess, for instance.

The routine reporting of uncertainties has often been forgotten, particularly in demonstrations of automated detection systems such as for sea minehunting with high-frequency sonar (the author's field of research), where statistical rigor is likely to be upstaged by the novelty of a new algorithm design and its many details. At best one finds cautious conclusions regarding performance, with a warning that further testing is required to be conclusive, but with no indication of how much more testing would be required to be decisive. At worst, one finds the claim of "perfect performance" (all targets detected, with no false alarms) when the small number of dummy targets used in the test provides little confidence. Indeed, perfect test results occur often enough in practice owing to small sample sizes, and warrant special attention here. The same tendencies are encountered with military detection systems more generally, whenever the immediate questions of understanding and demonstrating new systems or technology overshadow the larger questions of overall operational performance.

There are several well-established, traditional approaches to managing uncertainties, of which two are considered here: hypothesis testing and Bayesian inference. The two methods are complimentary, and their end results are much the same as we shall see. Neither method is decidedly better (simpler or more immediately evident) than the other, but together they show that there are no shortcuts to pass-fail performance testing; that some degree of statistical complexity is unavoidable.

The motivation behind this work has been two practical questions that invariably arise in performance testing:

1. How many dummy targets must be deployed, and how many clutter events must be encountered, in order to prove that a particular detection system meets a given performance level?
2. With what confidence can an objective observer claim that the detection system under test exceeds given performance levels?

Both are addressed here. Dummy targets and clutter events are mentioned in the first because both target and clutter classes are required in a complete measure of performance under the Neyman-Pearson criteria for detection systems, which includes the probability of detection and false alarm [1] [2]. To omit one is to leave the performance in doubt inasmuch as the detection thresholds of any detector can be adjusted for perfect performance in one respect at the expense of the other—i.e., perfect detection of all targets with intolerably many false alarms, or perfect rejection of all clutter with poor detection of targets. The quality of a detector therefore lies in the balance it strikes between the two classes, not in its excellence against one class alone.

Our approach will be to first give, in Section (2), an example of performance uncertainties, and to clarify the scope and context of the present work by defining terms and reviewing the binomial probability distribution which governs performance testing. A formal analysis of performance measurement uncertainties follows in Section (3). Pass-fail decisions are taken up in Section (4) using hypothesis tests, and then again in Section (5) using Bayesian inference. A treatment of the perfect test result and an estimate of the minimum number of dummy targets is included in each of those sections.

2. Preliminary remarks about detection performance testing

2.1 Demonstration of uncertainties

Fig.(1) illustrates the uncertainties associated with pass-fail performance tests. It shows first of all the minimum allowable receiver-operator characteristics (ROC) [1], separating the "pass" region (above and left of the curve, signifying higher probability of detection for a given probability of false alarm) from the "fail" region (below and right of the curve). The dashed line represents the ROC for the hypothetical system under test, which would not be known in advance of a test, of course, but is assumed here for the demonstration of uncertainties only. Finally, the individual points represent 300 independent performance tests, each computed here by statistical simulation, using the binomial distribution given in Section (2.6), for randomly chosen operating points on the ROC curve of the system under test. It was assumed that 10 dummy targets and 100 clutter events were used in the computation of each point. The scatter of the points illustrates the random variability between repeated tests, due solely to the fact that we are estimating the performance probabilities on the basis of a finite target and clutter sample sizes. In practice, we would of course have just one performance test, and no

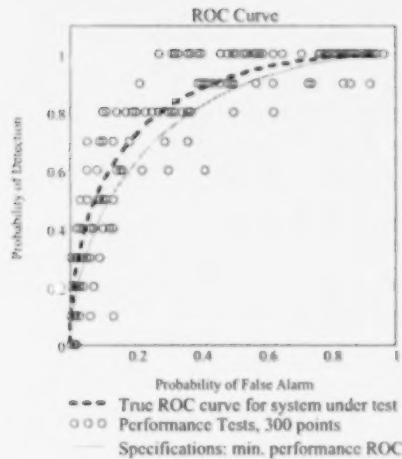


Figure 1: A demonstration of pass-fail test uncertainties: The points represent 300 independent pass-fail performance tests of a given detection system (dashed curve), relative to minimum performance specifications (solid curve), using 10 dummy mines and 100 clutter events in each test.

experimental indication of the variance through repeated trials as shown in the figure. It is clear that a pass-fail conclusion would be inconsistent on repeated trials, and that the conclusions drawn from a single test remain in question. Increasing the sample sizes reduces the variance, as we will see, and therefore increases the confidence in a pass-fail conclusion, but when is the sample size adequate to rule decisively, for either pass or fail?

2.2 Single-scenario test

A performance test is only useful insofar as we can draw objective inferences regarding performance in the wider field of real-world operations. The test must therefore be representative of operations in two respects:

1. the performance-affecting factors during the test must be the same as those in the field; and
2. the targets and clutter encountered during the test must be representative of those encountered in the field.

Failure in either case undermines the confidence placed in the conclusions drawn from the test.

The correspondence between the test and the intended operation will be easy or difficult to ensure, depending on the scope of the conclusions to be drawn. If the conclusions are to apply to one narrowly defined but realistic scenario (a single operational area or seafloor type, for instance), then the test could be carried out in just that scenario, by first deploying dummy targets, and then using the detection system under test to search for them, much as during actual operations. On the other hand, if the conclusions drawn from the test are to apply to all operational scenarios that could conceivably be encountered over the life of the detection system, then the tests must be carried out for all likely scenarios (assuming that all can in fact be identified in advance), and the results of those test must be combined in proportion to the prior probabilities of encountering each of those scenarios. The generalization of performance from a single to multiple scenarios has been considered in Section 3.1 of reference [3], though without an analysis of uncertainties. This would require a very ambitious performance testing program. Typically one would of necessity focus on a few most likely, or immediately applicable scenarios for testing. The single-scenario test is the essential component in any case, and is assumed throughout this paper.

2.3 Clutter events

The term "clutter events" in connection with false alarms (question 1, page 2, and throughout this paper) can be difficult to define precisely. They are the set of independent non-target signatures to which we assign a probability of triggering a false alarm. Membership in this set may depend in part on the detector itself. A detector based on energy detection, for instance, in which a detection is registered when the signal intensity exceeds a given threshold [4] [5], may require that every pixel in the signal constitutes a clutter event, in which case the total number of clutter events is likely to be extremely large, and the required probability of false alarm proportionally small. For a matched filter or correlation detector [6], the number of clutter events might be estimated by the ratio of the total signal length (or area) divided by the filter's sliding window size. Then again, for a detector designed to discriminate very particularly between similar objects, such as mines and mine-sized rocks in sidescan sonar imagery [7] for instance, the number of clutter events would be the number of rocks encountered during the test. In any case, it is important to clearly define clutter events for a performance test in order to evaluate a detector's false alarm performance.

2.4 Performance specifications

Pass-fail testing begins with a specification of the minimum allowable performance. To be clear about these specifications, we begin as in elementary detection theory, by assuming that the world, as seen by the detector, can be divided into two classes: non-targets, or clutter (class 0), and targets (class 1). Let $\overline{p}_0$ be the conditional probability of registering a false alarm when given a member of class 0, and $\overline{p}_1$ be the conditional probability of registering a correct detection given a member of class 1. These constitute the performance probabilities under the Neyman-Pearson design

criterion for detectors. The overhead bars signify that these parameters represent the "true" performance that one would like to measure with certainty, but for which the performance test only provides estimates.

A pass-fail performance test undertakes to determine whether the true performance exceeds minimum requirements,

$$\begin{aligned}\overline{p_0} &< s_0 \text{ and} \\ \overline{p_1} &> s_1;\end{aligned}\tag{1}$$

s_0 being the maximum allowable probability of false alarm, and s_1 the minimum allowable probability of detection. These inequalities must be accepted for the system to pass.

Let p_0 be the ratio of the number of false alarms to the total number of clutter events in the test, and p_1 be the ratio of the number of targets correctly detected to the total number of targets deployed. The true performance probabilities are the expected values each

$$\overline{p_i} = E\{p_i\},\tag{2}$$

where $i = 0$ or 1 to treat both classes at once. A performance test is too often viewed as a test of the inequalities, with the estimates p_i substituted for the true performance $\overline{p_i}$,

$$\begin{aligned}p_1 &> s_1 \text{ and} \\ p_0 &< s_0.\end{aligned}\tag{3}$$

Being random variables, the estimates p_i are subject to random variations, making the conclusions unrepresentative of the true performance. A pass might be the result of "good luck" on a single trial, or a failure might be the result of "bad luck". This uncertainty is widely recognized, and accounts for the cautious qualifications that researchers typically append to performance tests, that more testing is needed to be decisive, but a statistical treatment of uncertainty is required to be objectively and finally decisive.

2.5 The marginal detection system

The shortcomings of (3) become most apparent if we imagine that we happen to be evaluating a *marginal* detection system—i.e., one whose true performance $\overline{p_i}$ lies close to the specifications, $\overline{p_i} \approx s_i$. Because the true performance lies near the pass-fail decision boundaries, we might expect that, owing to uncertainties in the test, the system would pass the detection ($i = 1$) part of the test (3) with a probability of about 0.50, and therefore fail with roughly the same probability¹. Likewise for the clutter rejection ($i = 0$) part of the test. Hence the marginal detection system would be expected to fail,

¹This assumes probability densities that are symmetric about the mean; hence a 0.5 pass-fail probability for the detection and false alarm tests independently. The symmetric assumption does not strictly apply here, but serves for illustration only.

in at least one respect (false alarms or detections), in roughly 75 % of the performance tests to which the system is subjected. The marginal detection system is in fact the most difficult to conclusively evaluate for this reason.

2.6 Measured performance: a binomial random variable

Let the number of targets and clutter events in a trial be n_1 and n_0 , respectively, and the measured number of detections and false alarms be k_1 and k_0 . The number of detections k_i registered in n_i independent samples follows a binomial random distribution [9], such that the probability of registering k_i detections in a test given the true system performance $\bar{p}_i$, is

$$P(k_i | \bar{p}_i) = \binom{n_i}{k_i} \bar{p}_i^{k_i} (1 - \bar{p}_i)^{n_i - k_i}; \quad (4)$$

In practice, the number of dummy mine targets n_1 for detection is not large, much less than 200 typically, and $P(k_1)$ can be computed straightforwardly using (4). But if the number of clutter events n_0 is very large, and p_0 is small, then the binomial distribution is best approximated by Poisson's distribution [8]

$$P_0(k_0 | \bar{p}_0) = \frac{(n_0 \bar{p}_0)^{k_0} e^{-n_0 \bar{p}_0}}{\Gamma(k_0 + 1)}, \quad (5)$$

in which Γ is the Gamma function

$$\Gamma(k + 1) = k!. \quad (6)$$

The expected detection count in either case is

$$E\{k_i\} = n_i \bar{p}_i, \quad (7)$$

and the standard deviation is [9]

$$\sigma_{k_i} = \sqrt{n_i \bar{p}_i (1 - \bar{p}_i)}. \quad (8)$$

The measured performance is the proportion

$$p_i = \frac{k_i}{n_i}, \quad (9)$$

whose standard deviation, by (8), is

$$\sigma_{p_i} = \sqrt{\frac{\bar{p}_i (1 - \bar{p}_i)}{n_i}}. \quad (10)$$

The uncertainty σ_{p_i} of the measured performance p_i clearly decreases as the number of samples n_i increases. The true performance $\bar{p}_i$ is not known in advance, of course, so another approach to uncertainties is required.

3. Uncertainty analysis

The common practice of expressing uncertainties in terms of the standard deviation [8], [9] is based on the assumption that the measurement is a Gaussian random variable, distributed symmetrically about its mean. But the binomial distribution is not symmetric, especially for high performance detectors ($p_0 \rightarrow 0, p_1 \rightarrow 1$), and its standard deviation is furthermore dependent on the mean, all of which makes the standard deviation of little use for directly assessing uncertainties in performance measurements. The essential approach to uncertainties is nevertheless the same as it is with the Gaussian; that is, by first of all defining a *confidence interval*, or error bounds, for the measurement—i.e., the bounds within which the true value of a parameter lies with a given *confidence probability* close to unity. In this section we review the confidence intervals for performance tests following the method of Kendall and Stuart [10, Section 20.9].

3.1 Confidence intervals

Let the lower and upper confidence bounds constraining the true system performance be called $\bar{t}_i^A$ and $\bar{t}_i^B$, respectively. The probability that the true system performance $\bar{p}_i$ lies within the confidence interval, given the measured performance p_i can be written as

$$P(\bar{t}_i^A < \bar{p}_i < \bar{t}_i^B | p_i) = c_i, \quad (11)$$

in which c_i is the confidence probability. To determine the interval $\bar{t}_i^A$ to $\bar{t}_i^B$, we begin by first computing a similar interval for the more straightforward inverse of (11),

$$P(t_i^A < p_i < t_i^B | \bar{p}_i) \geq c_i; \quad (12)$$

that is, the probability P that the measured performance p_i will lie within the interval t_i^A to t_i^B , given the true performance $\bar{p}_i$. The inequality is necessary because the measured performance p_i , t_i^A , and t_i^B belong to the set of rational numbers as in (9), which permits equality almost nowhere in the range $t_i^A < p_i < t_i^B$. We rewrite (12) as

$$P(k_i^A < k_i < k_i^B | \bar{p}_i) \geq c_i, \quad (13)$$

where k_i is defined as in (9), and $k_i^{A,B}$ are the discrete counterparts of $t_i^{A,B}$,

$$\begin{aligned} k_i^A &= \text{floor}(n_i t_i^A), \\ k_i^B &= \text{ceil}(n_i t_i^B); \end{aligned} \quad (14)$$

the floor (x) operation returning the largest integer less than x , and ceil (x) returning the smallest integer bigger than x .

Equation (13) constrains width of the interval $k_i^B - k_i^A$ given one of its end points, but not the value of each bound independently. To determine k_i^A and k_i^B , it is therefore

customary (as with Gaussian confidence intervals) to choose *central* intervals; that is,

$$\begin{aligned}\alpha_i/2 &\geq P(k_i < k_i^A | \bar{p}_i); \\ \alpha_i/2 &\geq P(k_i > k_i^B | \bar{p}_i); \end{aligned} \quad (15)$$

in which the *significance*

$$\alpha_i = 1 - c_i, \quad (16)$$

has been introduced for use in hypothesis testing in the next section. For the moment, however, inserting the binomial distribution into (4), we have

$$\begin{aligned}\alpha_i/2 &\geq \sum_{k=0}^{k_i^A} P(k | \bar{p}_i); \\ \alpha_i/2 &\geq \sum_{k=k_i^B}^{n_i} P(k | \bar{p}_i); \end{aligned} \quad (17)$$

the first of which determines the lower bound k_i^A , and hence $t_i^A = k_i^A/n_i$; the second determines the upper bound k_i^B , and hence $t_i^B = k_i^B/n_i$. In practice, k_i^A and k_i^B must be determined numerically.

This has been done for detection performance in Fig.(??), for example, for three different confidence probabilities. Note that the figure has been constructed *horizontally*, with the independent variable, the true system performance $\bar{p}_i$, on the vertical axis, and with the dependent variable, the measure performance p_i and its corresponding confidence interval t_i^A and t_i^B , along the horizontal axis². But what we would like to do is to read the graph *vertically*—i.e., given a measured value of p_i , read points (p_i, t_i^A) and (p_i, t_i^B) along a vertical line through p_i , to get the confidence interval t_i^A and t_i^B bounding the true performance $\bar{p}_i$, as intended from the outset (11). The justification for reading the graph vertically follows by noting first of all that each test of an actual system is characterized by a point $(p_i, \bar{p}_i)$ somewhere on the graph—the first coordinate p_i being known with certainty because it is the measured result, and the second $\bar{p}_i$ remaining unknown because it is the true system performance. And, secondly, noting that if the point $(p_i, \bar{p}_i)$ falls in the interval, then it does so with a probability equal to or greater than c_i , for that was the condition on which the interval was originally designed. Thus, having measured the proportion p_i detected out of a total sample population n_i for the test, we can read its corresponding confidence interval by looking vertically along the line p_i , from the lower bound t_i^A to the upper t_i^B .

Note that the interval in Fig.(??) widens, and hence, the uncertainty regarding to true system performance increases, as the confidence probability is increased, signifying

²Note that the confidence intervals are central intervals provided that $t_i^A \neq 0$ and $t_i^B \neq 1$. This is not strictly true at the extreme left or right sides of the graph where $t_i^A = 0$ or $t_i^B = 1$, and where strong asymmetry in the binomial distribution makes the central interval impossible. The confidence probability therefore varies slightly in the horizontal, near the extreme ends of the graph where $t_i^A = 0$ or $t_i^B = 1$; varies, that is, between c_i at points $t_i^A \neq 0$ and $t_i^B \neq 1$, and $c_i + \alpha_i/2$ at the extreme ends $(p_i, \bar{p}_i) = (0, 0)$ and $(p_i, \bar{p}_i) = (1, 1)$.

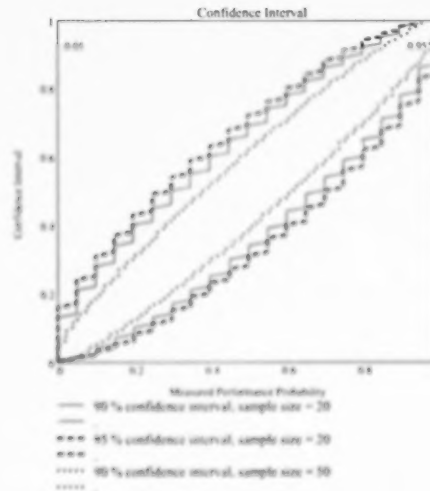


Figure 2: Three confidence intervals are plotted here as a function of the measured performance, assuming that the sample size is 20. If the measured probability of detection were $P_1 = 0.80$ (on the horizontal axis), for instance, then, using the solid lines, the probability that the true system performance lies between 0.65 and 0.95 (vertical axis) is greater than 90 %. The three intervals show how the interval widens as the confidence probability is increased (dashed lines relative to solid), but narrows as the sample size is increased (dotted lines relative to solid). The vertical lines near the extreme edges of the graph mark the points at which the asymmetry in the binomial distribution makes central intervals impossible, as noted in an earlier footnote.

evenmore stringent pass-fail criterion. On the other hand, the interval narrows, and hence, then uncertainty decreases, as the sample size is increased. The sample size must therefore be large enough to achieve the desired confidence and certainty at once.

4. Hypothesis testing

The performance specifications (1) are hypotheses that we would like to prove or disprove, accept or reject, for a given detection system. In statistics, the *rejection* of a hypothesis means that the hypothesis is contradicted by the evidence, whereas *acceptance* plays a weaker role, meaning no more than that the hypothesis is not contradicted by the evidence [9]. Acceptance is never a positive assertion that the hypothesis is true. If our performance test is to be objectively conclusive, we must recast our hypotheses accordingly, by posing a *null hypotheses* that the test is designed to reject.

To this end, recall that the *marginal* detection system is the most difficult to pass or fail conclusively. If the performance test is to be in any way conclusive, then we must first

prove that the detection system is not performing marginally by *rejecting* the hypotheses, H_i

$$H_i : p_i = s_i. \quad (18)$$

A performance test remains inconclusive for lack of evidence whenever H_i cannot be rejected on the basis of its observed performance, and more testing is required.

We therefore define an *acceptance* zone for the measured performance, $t_i^A < s_i < t_i^B$, such that the hypotheses H_i are accepted if their respective frequencies p_i fall inside that zone:

$$t_i^A < p_i < t_i^B \implies \text{accept } H_i. \quad (19)$$

The regions outside those bounds are called the *critical zone*, in which the marginal detection system can be confidently ruled out (H_i rejected), leaving us in a position to draw pass-fail inferences about its performance:

$$\begin{aligned} p_1 &> t_1^B && \text{for detection pass,} \\ p_1 &< t_1^A && \text{for detection fail,} \\ p_0 &< t_0^A && \text{for false alarm pass,} \\ p_0 &> t_0^B && \text{for false alarm fail} \end{aligned} \quad (20)$$

It is in the selection of the decision boundaries $t_i^{A,B}$ that the confidence placed in the pass-fail conclusions is formulated objectively. To this end, let α_i be the maximum probability that we shall by prior agreement allow of mistakenly rejecting H_i when it is in fact true,

$$\alpha_i > P \{ \text{rejecting } H_i \mid H_i \text{ is true} \}, \quad (21)$$

formally called a Type I error [9]. α_i is called the *significance* of the hypothesis test. It is the probability of mistakenly passing or failing the marginal detection system, on the basis of the relative frequency p_i , when the sample size n_i is in fact too small to substantiate such a conclusion. Choosing smaller α_i makes the pass-fail conclusions drawn from the test more confident. One traditionally chooses $\alpha_i = 0.10$ or 0.05 .

Equation (21) uniquely determines the decision boundaries $t_i^{A,B}$ if we apportion the probability of making a Type I error equally among the possible errors:

$$\begin{aligned} \frac{\alpha_1}{2} &> P \{ \text{rejecting } H_1 \text{ because } p_1 > t_1^B \mid H_1 \text{ is true} \}, \\ \frac{\alpha_1}{2} &> P \{ \text{rejecting } H_1 \text{ because } p_1 < t_1^A \mid H_1 \text{ is true} \}, \end{aligned} \quad (22)$$

$$\begin{aligned} \frac{\alpha_0}{2} &> P \{ \text{rejecting } H_0 \text{ because } p_0 < t_0^A \mid H_0 \text{ is true} \}, \\ \frac{\alpha_0}{2} &> P \{ \text{rejecting } H_0 \text{ because } p_0 > t_0^B \mid H_0 \text{ is true} \}. \end{aligned} \quad (23)$$

These can be written as

$$\mathbb{Q}_1^A > \sum_{k=k_1^A}^{n_1} P_1(k | \bar{p}_1 = s_1), \quad \mathbb{Q}_2^A > \sum_{k=0}^{k_2^A} P_0(k | \bar{p}_0 = s_0), \quad (24)$$

and

$$\mathbb{Q}_1^B > \sum_{k=0}^{k_1^B} P_1(k | \bar{p}_1 = s_1), \quad \mathbb{Q}_2^B > \sum_{k=k_2^B}^{n_2} P_0(k | \bar{p}_0 = s_0), \quad (25)$$

respectively, in which $k_i^{A,B}$ are the discrete, alarm-count decision boundaries corresponding to the original decision boundaries $t_i^{A,B}$ as in (14).

Much as with the confidence intervals in the previous section, the decision boundaries $k_i^{A,B}$ and $t_i^{A,B}$ can be determined numerically from (24) and (25), using the appropriate binomial distribution (4) or (5) for P_i , together with solution technique akin to root finding. This process is illustrated graphically in Figure (3), which plots the right side of equation (24) and (25) for detection (class $i = 1$) along the vertical axis, as a function of trial threshold values along the horizontal axis. The threshold values at which these conditionals cross the significance level $\alpha/2$ of the two-tailed test are the threshold values that determine the critical and acceptance zones, t_1^A and t_1^B in (19), by which the performance is to be judged.

Figure (4) shows the upper and lower thresholds as a function of the detection performance specification s_1 , for sample sizes of $n_1 = 10, 20$ and 50 dummy targets, and significance $\alpha_1 = 0.10$. The upper and lower thresholds approach each other as the sample size increases, indicating that the test is more likely to be conclusive as the sample size is increased, as we would expect.

To illustrate the use of Figure (4), let us assume that the minimum required detection performance for the system under test is $s_1 = 0.80$. Finding this point along the horizontal axis, then looking upwards to the lower and upper thresholds for $n_1 = 20$ say, we see that the upper and lower decision thresholds for a pass-fail test are $t_1^A = 0.65$ and $t_1^B = 0.95$, respectively. If the measured performance p_1 in (9) falls below t_1^A , then the system is judged to have failed with respect to the detection specification. If it falls above t_1^B , then the system is judged to have passed. If it falls between t_1^A and t_1^B , then the test remains inconclusive because the null hypothesis (18) cannot be ruled out; the sample size n_1 is too small. Presumably one could extend the analysis to estimate the fewest number of additional targets that would be required to make a pass-fail judgment with the desired significance, but that is beyond our present scope.

Figures such as (3) and (4) could be constructed for any significance α_1 and sample size n_1 . Similar graphs could be for false alarm performance testing as well.

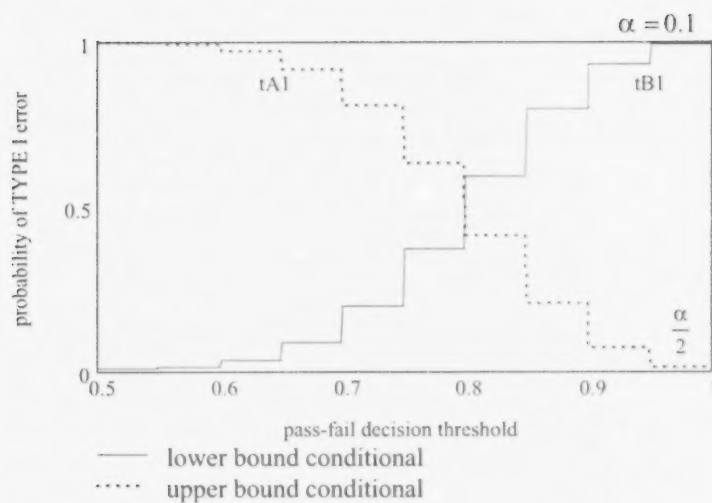


Figure 3: The pass-fail decision thresholds can be determined graphically. A detection (class 1) performance test is assumed here. The sample size for the figure is $n_1 = 20$, and the performance specification is $s_1 = 0.80$. The horizontal axis represents trial threshold values, and the vertical axis is the probability of making an upper or lower TYPE 1 error. The threshold values at which these equal $\alpha/2$ are the pass-fail threshold values for the test. If the measured performance falls below t_1^A , then the system fails to meet the detection specification s_1 . If it falls above t_1^B , then the system passes. If it falls between the two thresholds, then more testing (a larger sample size) is required.

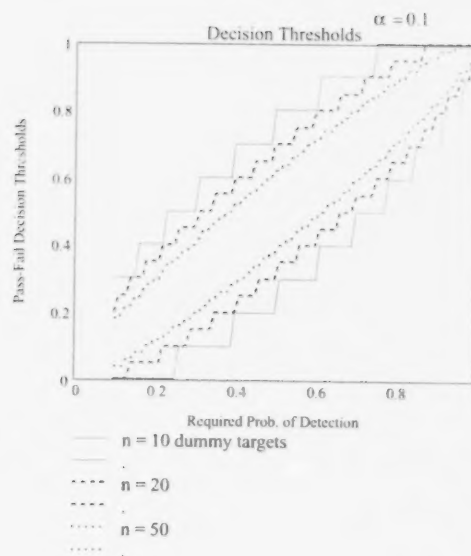


Figure 4: The pass-fail decision boundaries in Figure (1) can also be determined numerically. This has been done here for a range of detection performance specifications s_1 (horizontal axis), and sample sizes $n_1 = 10, 20$ and 50 , assuming $\alpha_1 = 0.10$.

4.1 Interpreting the perfect result

If perfect test results are observed (i.e., all dummy targets detected, with no false alarms), then performance failure can be ruled out and the hypothesis test becomes one sided: either the system passes the performance specifications s_i , or the sample size is too small for the test to be significant. To decide between these alternatives we first solve (21) for the minimum detection performance specification s_1 that passes with the desired significance α_1 ,

$$\alpha_1 > \sum_{k=n_1}^{n_1} P_1(k | \bar{p}_1 = s_1) = (s_1)^{n_1}, \quad (26)$$

whereby

$$s_1 < \alpha_1^{1/n_1}. \quad (27)$$

Thus, on the basis of an unbiased performance test in which all of the n_1 deployed dummy targets are in fact detected, one can conclude (with significance α_1) that the system performance $\bar{p}_1$ exceeds the detection specification s_1 when the condition (27) is satisfied. If it is not satisfied, then the sample size n_1 is too small to be conclusive.

Solving for n_1 , the minimum number of dummy mines in any detection test is

$$n_1 > \frac{\log(\alpha_1)}{\log(s_1)}, \quad (28)$$

as plotted in Figure (5). To illustrate, sea minehunting systems may typically call for minimum probability of detection $s_1 = 0.80$, in which case $n_1 > 10.3$ for $\alpha_1 = 0.1$. In other words, at least 11 dummy mines (or an equivalent number of multiple deployments of fewer mines) are required in any detection performance test. Significantly more dummy mines would be required in practice, however, because this is a lower limit based on the assumption that all mines will be detected by the system in a random test.

Likewise, to make a pass decision in the case of false alarms, we solve (21) for the maximum false alarm performance specification s_0 that passes with the desired significance α_0 ,

$$\alpha_0 > \sum_{k=0}^0 P_0(k | \bar{p}_0 = s_0) = (1 - s_0)^{n_0}, \quad (29)$$

whereby

$$s_0 < 1 - \alpha_0^{1/n_0}. \quad (30)$$

then the detection system passes the false alarm performance specification with significance better than α_0 . Solving for n_0 , the minimum number of clutter events for a test of false alarms is therefore

$$n_0 > \frac{\log(\alpha_0)}{\log(1 - s_0)}. \quad (31)$$

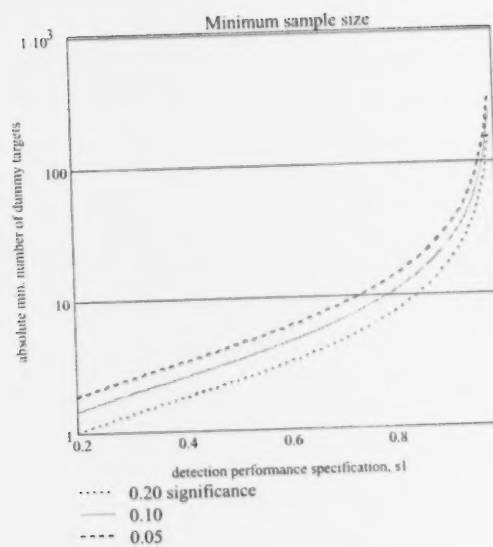


Figure 5: The minimum number of dummy targets for a performance test presumes, very optimistically, that all of the targets deployed in a random, unbiased test will be detected. In practice, significantly more targets would be required.

If clutter events were mine-like rocks, for instance, for which the maximum probability of false alarm was $s_0 = 0.05$, then the number of mine-like rocks required for the test is $n_0 > 44.89$ for $\alpha_0 = 0.1$. In other words, at least 45 different mine-like rocks are required in a test of false alarms. Here again, one would like to see significantly more clutter events in practice because this lower bound assumes that perfect clutter rejection will be observed during the test. Figure (5) can be applied for false-alarm tests, by simply making $1 - s_0$ the abscissa, and minimum n_0 as the ordinate.

5. Bayesian inference

The question of confidence can also be addressed using Bayesian inference. That is, given that the measured performance is p_i , one would like to know the probability that the system's actual performance $\bar{p}_i$ exceeds the specifications s_i . From Bayes theorem for inferred probabilities [11],

$$P(\bar{p}_i \text{ passes} | p_i) = \frac{P(p_i | \bar{p}_i \text{ passes}) P(\bar{p}_i \text{ passes})}{P(p_i)}, \quad (32)$$

where $P(\bar{p}_i \text{ passes})$ is the prior probability that the CAD system's performance exceeds the performance specification. Assuming the unbiased stance that any value of the system's actual performance $0 < \bar{p}_i < 1$ is equally likely (i.e., maximum prior uncertainty [12]), we have

$$\begin{aligned} P(\bar{p}_1 \text{ passes}) &= 1 - s_1; \\ P(\bar{p}_0 \text{ passes}) &= s_0. \end{aligned} \quad (33)$$

$P(p_i | \bar{p}_i \text{ passes})$ is the prior probability of getting the measured performance given that the system exceeds the performance specifications, which can be determined from the binomial distribution $P(p_1 | \bar{p}_1)$ in (4),

$$\begin{aligned} P(p_1 | \bar{p}_1 \text{ passes}) &= P(p_1 | \bar{p}_1 > s_1) = \frac{1}{1-s_1} \int_{s_1}^1 P(p_1 | \bar{p}_1) d\bar{p}_1; \\ P(p_0 | \bar{p}_0 \text{ passes}) &= P(p_0 | \bar{p}_0 < s_0) = \frac{1}{s_0} \int_0^{s_0} P(p_0 | \bar{p}_0) d\bar{p}_0. \end{aligned} \quad (34)$$

These can be evaluated numerically. They depend in part on the respective sample sizes n_i , by way of both the binomial distribution (4) and the set of rational fractions comprising the range of measured performances $p_i = k_i/n_i$.

Finally, the prior probability of measuring performance p_i is

$$P(p_i) = \int_0^1 P(p_i | \bar{p}_i) d\bar{p}_i. \quad (35)$$

Substituting (34) and (35) into (32) gives

$$P(\bar{p}_1 \text{ passes} | p_1) = \frac{\int_0^1 P(p_1 | \bar{p}_1) d\bar{p}_1}{\int_0^{s_1} P(p_1 | \bar{p}_1) d\bar{p}_1}, \quad (36)$$

$$P(\bar{p}_0 \text{ passes} | p_0) = \frac{\int_0^{s_0} P(p_0 | \bar{p}_0) d\bar{p}_0}{\int_0^1 P(p_0 | \bar{p}_0) d\bar{p}_0}. \quad (37)$$

These are the (posterior) probabilities that a system passes the performance specifications when detections and false alarms are considered independently, given

1. the relative frequencies $p_i = k_i/n_i$ measured in a performance test,
2. the sample size n_i ,
3. that the test was unbiased, and
4. that we assume maximum uncertainty regarding the outcome of the test.

The fail probability may be derived in much the same way, and turns out to be

$$P(\bar{p}_i \text{ fails} | p_i) = 1 - P(\bar{p}_i \text{ passes} | p_i). \quad (38)$$

To decide whether a system passes or fails the detection or false alarm specification, one would first of all compute the pass probabilities $P(\bar{p}_i \text{ passes} | p_i)$, and then compare them against a minimum allowable threshold P_B , presumably near unity, which represents the degree of confidence desired in the pass-fail conclusions drawn from the test,

$$\begin{aligned} P(\bar{p}_i \text{ pass} | p_i) &> P_B \text{ for a pass,} \\ P(\bar{p}_i \text{ fail} | p_i) &> P_B \text{ for a fail.} \end{aligned} \quad (39)$$

If neither condition is true, then a pass-fail decision must be reserved until further testing—i.e., the sample size n_i is increased—to make the test conclusive.

Figure (6) plots the pass probability as a function of the proportion of dummy targets detected p_1 when the minimum allowable detection performance for a pass is $s_1 = 0.80$, and when the total number of dummy targets used in the test is 10, 20, or 50. Note that the pass threshold (i.e., the point along the horizontal axis where the pass probability first exceeds c_1) approaches the minimum allowable detection performance s_1 as the sample size increases. This reiterates earlier remarks in Section (2.4), that the marginal system, whose performance is close to the required minimum, is the most difficult to pass or fail conclusively. Very large sample sizes will be required.

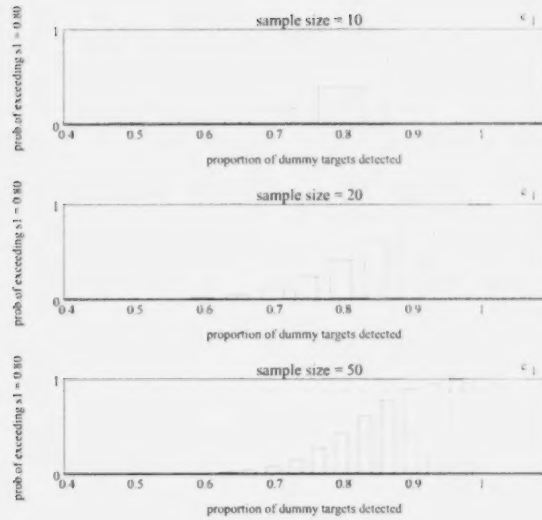


Figure 6: The pass probability is plotted here as a function of the measured detection performance, assuming sample sizes of 10, 20, and 50 dummy targets. The horizontal dashed line marks the $c_1 = 0.90$ pass probability. The system under test passes with regards to detection if the bar associated with the value of the performance measured in the test exceeds this level.

Figure (7) plots the pass decision threshold, above which the measured detection performance must fall for the system to pass, as a function of the minimum allowable detection performance s_1 for different confidence levels $P_B = 0.90, 0.95$, and 0.99 , assuming that there are 20 dummy mines in the test. This shows how the pass threshold becomes more stringent as the confidence required for a pass decision increases. Figure (8) follows much the same approach, but now with sample sizes of 10, 20 and 50 dummy targets, and fixed pass probability of 0.90. It shows how the pass threshold approaches the minimum allowable detection performance s_1 as the sample size increases.

5.1 Interpreting the perfect result

Substituting $p_1 = 1.0$ and $p_0 = 0$ into (36) and (37) gives

$$P_1 \leq P(\bar{p}_1 > s_1 \mid p_1 = 1) = \frac{\int_{s_1}^1 \bar{p}_1^{n_1} d\bar{p}_1}{\int_0^1 \bar{p}_1^{n_1} d\bar{p}_1} = 1 - s_1^{n_1+1} \quad (40)$$

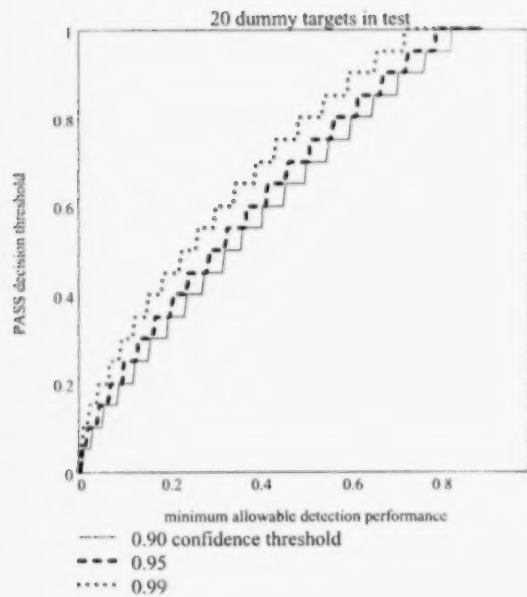


Figure 7: The system passes with respect to detection when the measured detection performance exceeds the PASS decision threshold, plotted here as a function of the performance requirement s_1 , assuming that there are $n_1 = 20$ dummy targets in the test. Ever more stringent thresholds result as the pass probability c_1 is increased.

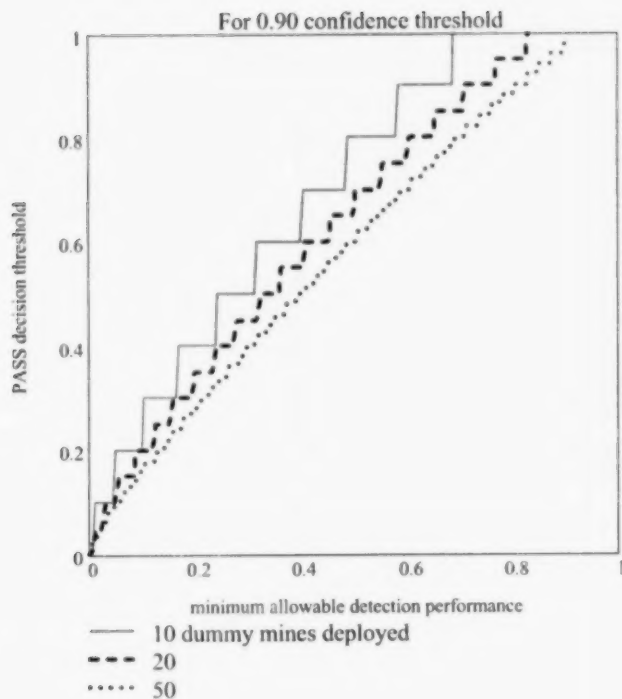


Figure 8: The system passes with respect to detection when the measured detection performance exceeds the PASS decision threshold, plotted here as a function of the performance requirement s_1 , assuming that the pass probability $c_1 = 0.90$, for three different sample sizes, $n_1 = 10, 20$ and 50 dummy targets. The pass threshold approaches the diagonal as the sample size increases.

and

$$P_0 \leq P(\bar{p}_0 < s_0 \mid p_0 = 0) = \frac{\int_0^{s_0} (1 - \bar{p}_0)^{n_0} d\bar{p}_0}{\int_0^1 (1 - \bar{p}_0)^{n_0} d\bar{p}_0} = 1 - (1 - s_0)^{n_0 + 1} \quad (41)$$

respectively. These are the (posterior) probabilities that a system exceeds the performance specifications (considered independently) when the perfect result is obtained. For example, if all of 5 dummy mines deployed in a trial were detected, and the detection performance specification to be exceeded were 80 %, then

$$P(\bar{p}_1 > 0.80 \mid p_1 = 1) \approx 0.74. \quad (42)$$

In other words, despite an apparent 100 % detection rate ($p_1 = 1.0$) found experimentally on a random sampling of 5 mines, the probability that the CAD system actually meets the $\bar{p}_1 > s_1 = 0.80$ performance specification is 74 %—notably less than absolute. The confidence increases as the sample size n_1 is increased. For $n_1 = 10$, for instance,

$$P(\bar{p}_1 > 0.80 \mid p_1 = 1) \approx 0.91. \quad (43)$$

Using (40) and (41) we can derive a lower bound for n_i , much as we did for hypothesis testing, but now using Bayesian inference,

$$n_1 > \frac{\log(1 - P_1)}{\log(s_1)} - 1, \quad (44)$$

and

$$n_0 > \frac{\log(1 - P_0)}{\log(1 - s_0)} - 1, \quad (45)$$

where P_i are the pass probabilities. It is interesting to note the similarity with the minimum sample size estimates (28) and (31) derived using the hypothesis test. For both class populations, n_0 and n_1 , the estimate using Bayesian inference is one less than the estimate using hypothesis testing if we assume that $\alpha_i = 1 - P_i$; the point being that the two approaches can be brought into good agreement for certain choices of significance α_i and pass probability P_i .

6. Conclusions

The need for proven system performance comes now to the forefront as detection systems evolve towards greater operational use. As with all experiments, measurements of detection performance are subject to uncertainties which can significantly undermine the confidence placed in any conclusions drawn from a measurement, such as whether

or not the detection system exceeds minimal performance specifications. The uncertainties have always been recognized by CAD system developers, but the estimation and management of those uncertainties have nevertheless been neglected.

Two statistical methods, hypothesis testing (Section (4)) and Bayesian inference (Section (5)), were applied to decide whether or not the system under test exceeds certain minimum performance requirements that end. Each method has its own measure of the confidence placed in a pass-fail decision. In hypothesis testing, confidence is quantified by the significance α —the probability of mistakenly drawing a pass-fail conclusions when the test has in fact been inadequate for the purpose (i.e., sample size was too small)—with smaller α signifying greater confidence. In Bayesian inference, confidence is quantified by the pass and fail probability threshold P_B —the threshold above which the pass and fail probabilities must fall for a pass or fail judgement to be asserted as true. The choice between the two methods, hypothesis test and Bayesian inference, is more a matter of preference than of relative advantages or disadvantages. Hypothesis testing may be more familiar because it is widely used elsewhere, but the significance α may be less intuitive than the Bayesian pass-fail probabilities. The numerical computations required for both methods are in many ways equivalent.

Special attention was given to drawing pass-fail conclusions when perfect performance is observed in a test (i.e., all targets detected; no false alarms), which occurs often in practice owing to the small sample sizes. The condition for passing a system on the basis of a perfect result was given in (27) and (30) for hypothesis testing, and in (40) and (41) for Bayesian inference. These permit the system evaluator to draw objective conclusions from perfect demonstrations, and prevent them from being swayed by unwarranted claims of high performance. The analysis led furthermore to estimates of the absolute minimum sample size needed for making pass-fail decisions (Fig.(5)), which is of use when planning performance tests. In practice, one would typically like to use significantly larger sample sizes than this minimum.

It is evident that all demonstrations of detection performance should at the very least report the number of targets deployed (n_1), the number of targets detected (k_1), the number of clutter events encountered (n_0), and the number of false alarms (k_0), because a thorough analysis of experimental uncertainties can be made from these four numbers in the manner shown here.

6.1 Future work

Here it was assumed that the minimum performance to be exceeded in a pass-fail test was specified by the minimum allowable detection performance (s_1) and a maximum allowable false alarm performance (s_0), that is to say, on the basis of just one point on its receiver-operator characteristic (ROC) curve. The pass-fail decision then separates into two independent tests, one for detection and one for false alarms, as we have seen here. Work is under way to generalize the specifications to a minimal allowable ROC curve, which represents the range of possible balances struck by a detection system

between the probabilities of detection and false alarm, by adjusting the detector's alarm thresholds.

References

1. Van Trees, H.L., *Detection, estimation, and modulation theory*, John Wiley and Sons, New York, 1968.
2. Koopman, B.O., *Search and Screening: general principles with historical applications*, Pergamon Press, New York, 1980.
3. Kessel R.T., "Mission-adaptive computer aided detection (CAD) for mine hunting: system concept and theory," Canadian National Defence Research and Development Branch Technical Memorandum DREA TM 2000-111, Feb 2001.
4. Paul, J., , Johnson, P., and Harris J., "*Final report for the Canadian remote minehunting sonar study*," Canadian National Defence Research and Development Branch Contractor Report DREA CR/96/451A, 1996.
5. A.D. Waite, *Sonar for practising engineers*, 2nd Ed, Thomson Marconi Sonar Ltd, Great Britain, 1998.
6. G.J. Dobeck, J.C. Hyland, and L. Smedly, "Automated detection/classification of sea mines in sonar imagery," Naval Research Reviews, Office of Naval Research, V.49, pp. 9-20, 1997.
7. Kessel, R.T., "Trainable linear filters for the automtated detection of mines in sidescan sonar images," Canadian National Defence Research and Development Branch Technical Memorandum DREA TM 1999-171, Dec 1999.
8. Shchigolev, B.M., *Mathematical Analysis of Observations*, Iliffe Books Ltd., London, 1965.
9. Walpole, R.E. and Myers, R.H., *Probability and Statistics for Engineers and Scientists*, 2nd Edition, MacMillan Publishing, New York, 1978.
10. Kendal, M., and Stuart, A., *The Advanced Theory of Statistics*, Charles Griffon and Co., 1979.
11. Papoulis, A., *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill Book Co., 1965.
12. Jeffreys, H., *Theory of probability*, 2nd Ed., Oxford, Clarendon Press, 1948.

UNCLASSIFIED
SECURITY CLASSIFICATION OF FORM
(highest classification of Title, Abstract, Keywords)

DOCUMENT CONTROL DATA		
(Security classification of title, body of abstract and indexing annotation must be entered when the overall document is classified)		
1 ORIGINATOR (the name and address of the organization preparing the document. Organizations for whom the document was prepared, e.g. Establishment sponsoring a contractor's report, or tasking agency, are entered in section 8.) Defence Research Establishment Atlantic PO Box 1012 Dartmouth, Nova Scotia, Canada B2Y 3Z7	2 SECURITY CLASSIFICATION (overall security classification of the document including special warning terms if applicable): UNCLASSIFIED	
3. TITLE (the complete document title as indicated on the title page. Its classification should be indicated by the appropriate abbreviation (S, C, R or U) in parentheses after the title). Pass-fail performance testing for detection systems		
4. AUTHORS (Last name, first name, middle initial. If military, show rank, e.g. Doe, Maj. John E.) Ronald T. Kessel		
5. DATE OF PUBLICATION (month and year of publication of document) Feb 2002	6a. NO OF PAGES (total containing information include Abstracts, Appendices, etc.) 23 (approx.)	6b. NO OF REFS (total cited in document) 12
7 DESCRIPTIVE NOTES (the category of the document, e.g. technical report, technical note or memorandum. If appropriate, enter the type of report, e.g. interim, progress, summary, annual or final. Give the inclusive dates when a specific reporting period is covered). Technical Memorandum (TM)		
8. SPONSORING ACTIVITY (the name of the department project office or laboratory sponsoring the research and development. Include address). Defence Research Establishment Atlantic		
9a. PROJECT OR GRANT NO. (if appropriate, the applicable research and development project or grant number under which the document was written. Please specify whether project or grant) 1da12	9b. CONTRACT NO. (if appropriate, the applicable number under which the document was written):	
10a. ORIGINATOR'S DOCUMENT NUMBER (the official document number by which the document is identified by the originating activity. This number must be unique to this document.) DREA TM 2001-205	10b. OTHER DOCUMENT NOs. (Any other numbers which may be assigned this document either by the originator or by the sponsor.)	
11. DOCUMENT AVAILABILITY (any limitations on further dissemination of the document, other than those imposed by security classification) (☒) Unlimited distribution () Defence departments and defence contractors; further distribution only as approved () Defence departments and Canadian defence contractors; further distribution only as approved () Government departments and agencies; further distribution only as approved () Defence departments; further distribution only as approved () Other (please specify):		
12. DOCUMENT ANNOUNCEMENT (any limitation to the bibliographic announcement of this document. This will normally correspond to the Document Availability (11). However, where further distribution (beyond the audience specified in (11) is possible, a wider announcement audience may be selected).		

UNCLASSIFIED
SECURITY CLASSIFICATION OF FORM

13. **ABSTRACT** (a brief and factual summary of the document. It may also appear elsewhere in the body of the document itself. It is highly desirable that the abstract of classified documents be unclassified. Each paragraph of the abstract shall begin with an indication of the security classification of the information in the paragraph (unless the document itself is unclassified) represented as (S), (C), (R), or (U). It is not necessary to include here abstracts in both official languages unless the text is bilingual).

Detection is an uncertain operation subject to many random factors. The performance of a detection system is therefore specified probabilistically, by way of its probabilities of detection and false alarm, and the evaluation of a system's performance falls, unavoidably, within the scope of probability and statistics. In military applications, the central role of probability and statistics has too often been upstaged by the novelty of new detection technology, which, to demonstrate in all of its features, typically leaves little time or inclination for a detailed treatment of performance probabilities. But a statistical analysis is crucial for drawing objective conclusions from a performance test, such as whether the system passes or fails minimum operational requirements. The statistics of performance testing are reviewed here, as a means to manage the measurement uncertainties in pass-fail system testing. Two different decision methods are presented, hypothesis testing and Bayesian inference, each with their particular approach to manage uncertainties, yet both working toward the same end. Pass-fail judgements drawn from "perfect" test results (no missed targets, and no false alarms) are considered because they are often encountered in practice owing to small sample sizes. The minimum number of dummy targets required for a performance test is derived, to serve as a rough guide when planning and evaluating performance demonstrations.

14. **KEYWORDS, DESCRIPTORS or IDENTIFIERS** (technically meaningful terms or short phrases that characterize a document and could be helpful in cataloguing the document. They should be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location may also be included. If possible keywords should be selected from a published thesaurus, e.g. Thesaurus of Engineering and Scientific Terms (TEST) and that thesaurus-identified. If it not possible to select indexing terms which are Unclassified, the classification of each should be indicated as with the title).

computer-aided detection, detection systems, automatic target recognition, minehunting, signal detection theory

UNCLASSIFIED
SECURITY CLASSIFICATION OF FORM

